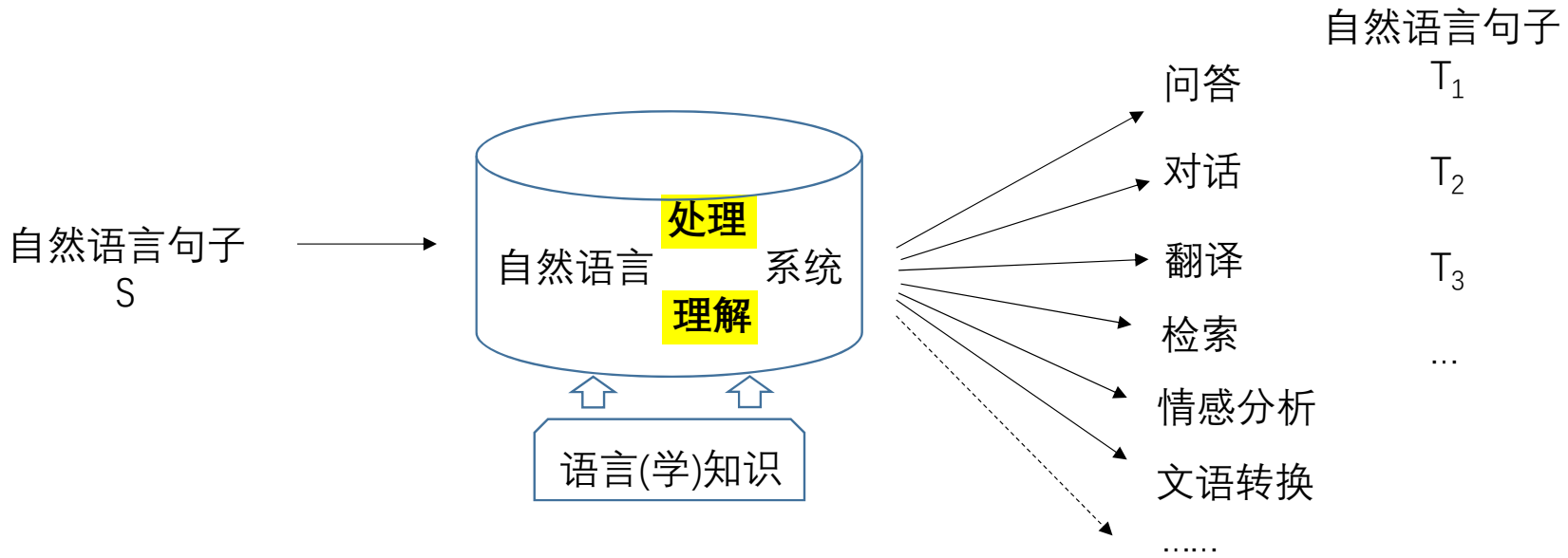
2018-11-02 跨学科交流会 北京大学

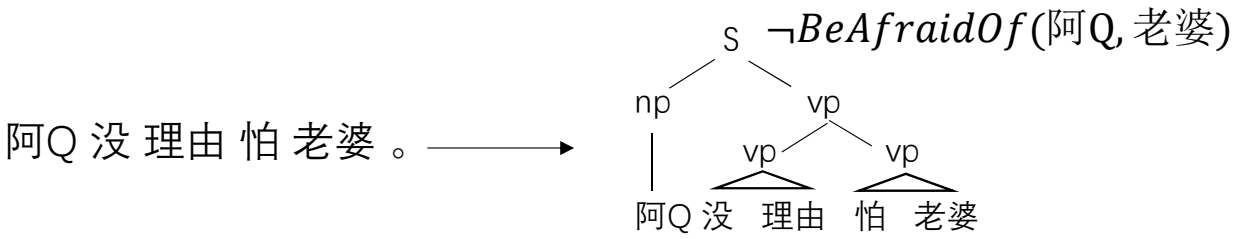
语言学知识：

—— 结构化、形式化、数据化、可视化

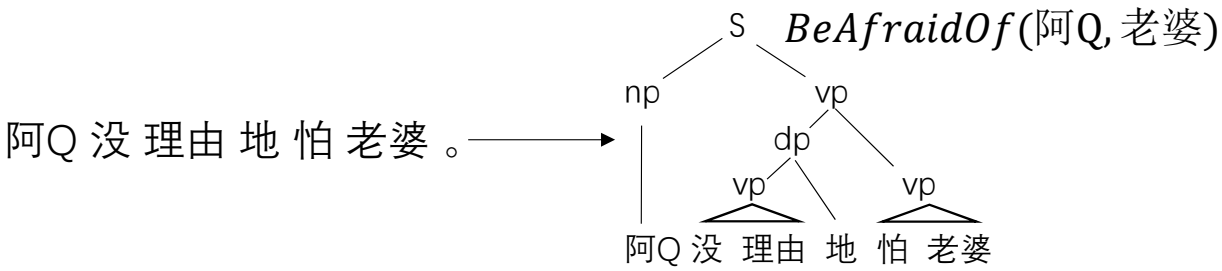
詹卫东
zwd@pku.edu.cn
北京大学中文系

- 面向计算的语言学研究的工作模式
- 对当前工作模式的反思 —— 路在何方?





Mr.Q has no reason to be afraid of his wife.



Mr.Q is afraid of his wife for no reason.

阿Q 没 理由 怕 老婆 。　Ah Q has no reason to be henpecked.

阿Q 没 理由 地 怕 老婆 。　A Q has no reason to be afraid of his wife.

黄金没理由涨。	There is no reason for gold to rise.
黄金没理由地涨。	Gold has risen for no reason.
她没理由害怕。	She had no reason to be afraid.
她没理由地害怕。	She had no reason to be afraid.
张老师从北京回来带给他们三个人每人一本书。	Teacher zhang came back from Beijing and brought a book to each of them.
张老师从北京回来带给他们三个人一人一本书。	Mr. Zhang came back from Beijing and brought them three people one book.
张老师从北京回来带给他们几个每人一本书。	Miss zhang came back from Beijing and brought them a book each.
他们三个人都读完了张老师要求今天必须完成的三篇文章。	All three of them have finished the three articles which teacher zhang required to finish today.
他们三个人都写完了张老师要求今天必须完成的三篇文章。	All three of them have finished the three articles which teacher zhang required to finish today.

http://fanyi.youdao.com/

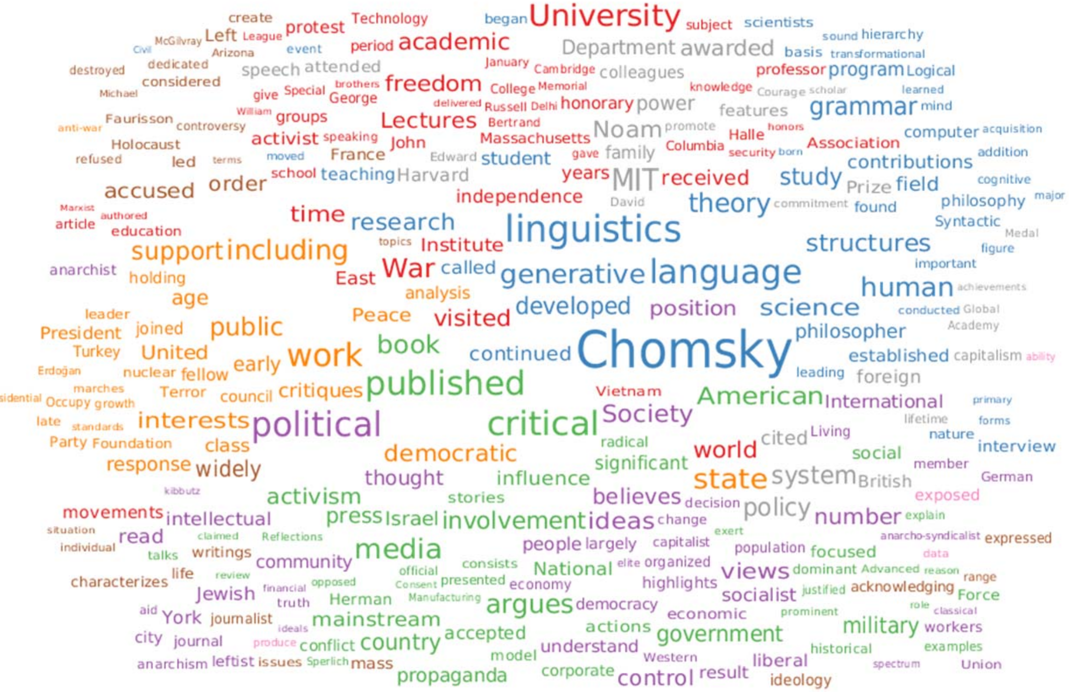
钱锋（1990）《计算语言学引论》，学林出版社1990年版。

语言学长啥样？ —— 语言学观念的嬗变

- 看作法律的语言学
- 看作生物学的语言学
- 看作化学的语言学
- 看作数学的语言学

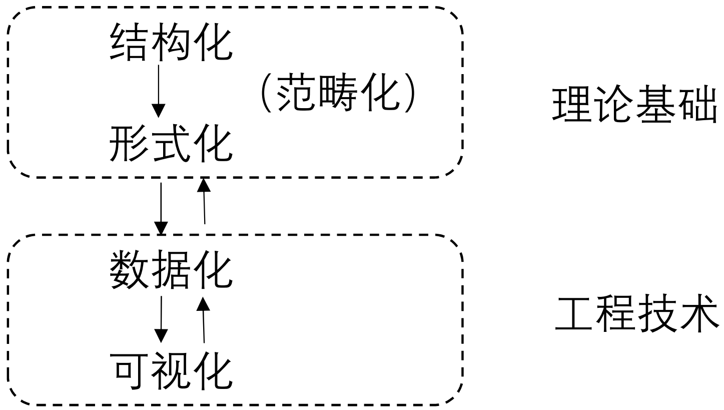
https://en.wikipedia.org/wiki/Noam_Chomsky

→ http://wordcloud.cs.arizona.edu/



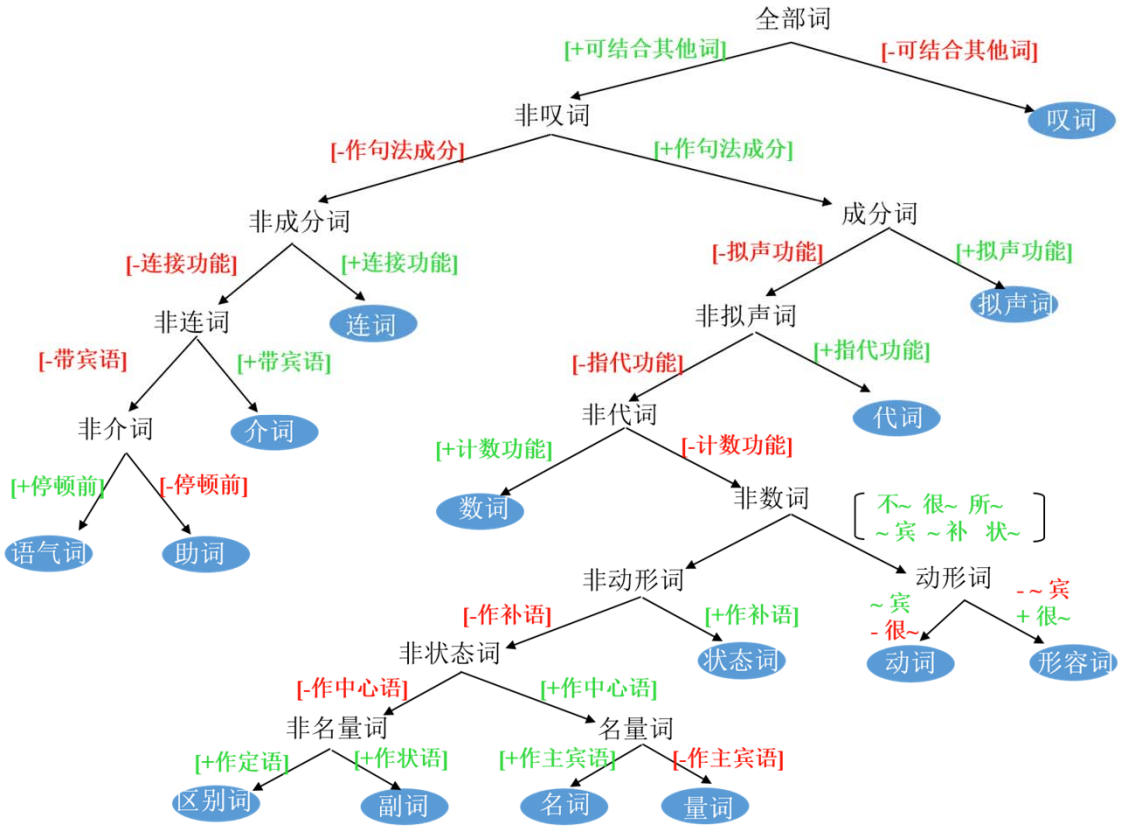
自然语言系统

- 显性：离散化、线性化
- 隐性：结构化、层次化



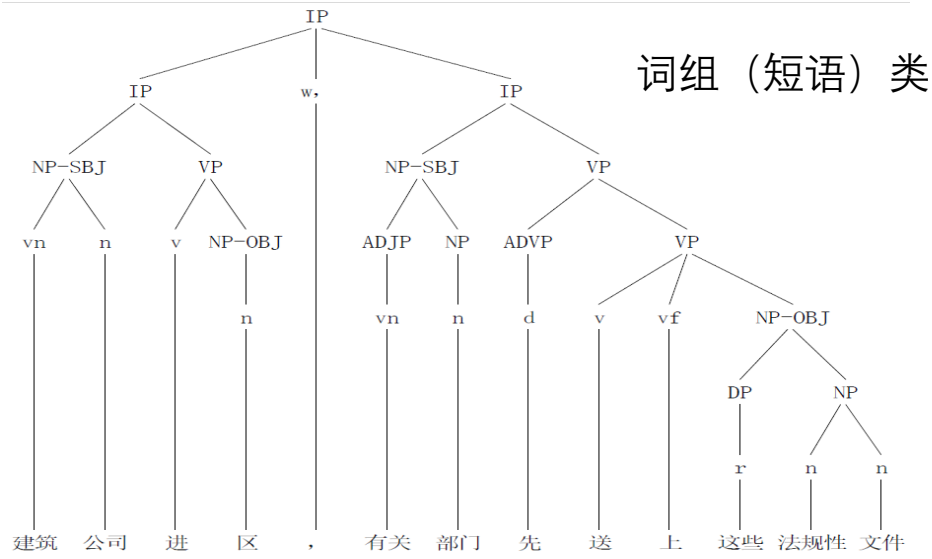
陆俭明 （2003）《现代汉语语法研究教程》§1.4，北京大学出版社

词类（范畴）　　词有多少类?

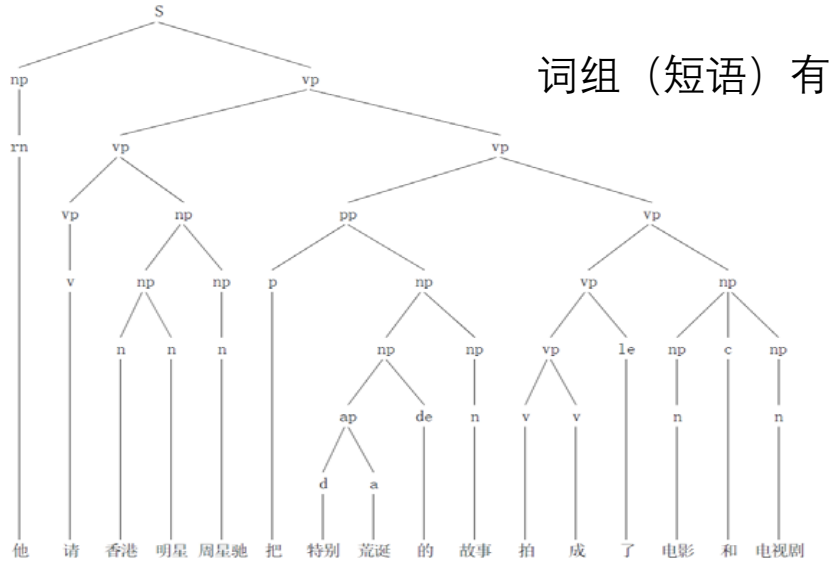


词语	词类	同形	义项	系词	助动	趋向	体谓准	双宾	单作补	复数主	后名	很	着了过	重叠	离合	兼类
保存	v						体				可		着了过	ABAB		
成为	v			系			体									
得到	v						体准						了过			
告诉	v						体谓	双					了过			
协商	v						体谓			复	可		了			
加以	v						准									
冒险	v										可		过	VVO	离	a
去	v	A1	除掉				体						了过	VV		
去	v	A2	~上海			趋	体		可				了过	VV		
去	v	B	扮演				体						了过			
应	v	A	答应						可				了			
应	v	B	应该		助		谓									
支持	v	1	支撑				体						着了过			
支持	v	2	鼓励并帮助				体谓准					很	着了过	ABAB		
指挥	v						体谓				可		着了过	ABAB		n
……																

俞士汶 等 1998《现代汉语语法信息词典详解》，清华大学出版社



词组（短语）类（范畴）



词组（短语）有多少类?

美国宾州大学汉语句法结构树库

IP → IP w IP
IP → NP VP
NP → vn n
NP → ADJP NP
NP → DP NP
NP → n n
VP → v NP
VP → ADVP VP
VP → v vf NP
……

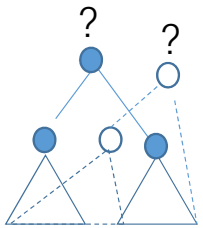
汉语句法结构形式规则

规则 该 长 啥样?
规则 该 有 多少?

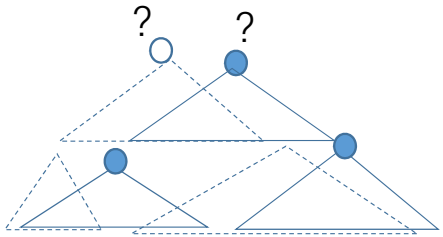
北京大学汉语句法结构树库

S → np vp
vp → vp vp
vp → vp np
vp → pp vp
vp → vp le
np → np np
np → ap de
np → np c np
pp → p np
……

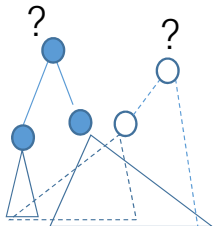
规则 该 长 啥样?



他 这 眼睛 是 看 书 看 的



老张 带 给 我们 每人 一本书



我 胖 我 的， 关 你 什么 事

规则 该 有 多少?

vp 规则: 232 条 （北大树库）

序号	规则	频次	示例
1496	vp -> !vp np	65756	vp(!vp(有)np(天大的困难))
1497	vp -> dp !vp	50316	vp(dp(也)!vp(要把这种产品试制成功))
1498	vp -> !vp vp	25309	vp(!vp(要)vp(把这种产品试制成功))
1499	vp -> pp !vp	17960	vp(pp(把这种产品)!vp(试制成功))
1500	vp -> !vp ule	11397	vp(!vp(进行)ule(了))
1501	vp -> vp !vp	10635	vp(vp(去办公室)!vp(开会))
1502	vp -> !v v	9024	vp(!v(传授)v(给))
1503	vp -> !vp sp	4996	vp(!vp(安装在)sp(桌子上))
1504	vp -> c !vp	4975	vp(c(就是)!vp(有天大的困难))
1505	vp -> !vp dj	4948	vp(!vp(告诉孩子)dj(他们一起去公园))
1506	vp -> !vp wco vp	4836	vp(!vp(给你一个桃子)wco(，)vp(接着))
1507	vp -> !v uzhe	4248	vp(!v(放)uzhe(着))
1508	vp -> !vp qp	3607	vp(!vp(分配)qp(一下))
1509	vp -> ap !vp	3062	vp(ap(容易)!vp(生病))
1510	vp -> !vp ap	3002	vp(!vp(感到)ap(满意))
1511	vp -> !v p	2823	vp(!v(安装)p(在))
1512	vp -> tp !vp	1875	vp(tp(早上)!vp(锻炼))
1513	vp -> !vp y	1770	vp(!vp(再买一本新的)y(吧))
1514	vp -> vp wco !vp	1703	vp(vp(不按客观规律办事)wco(，)!vp(非失败不可)
1515	vp -> !v np vp	1641	vp(!v(使)np(工业的面貌)vp(发生了巨大的变化))
1516	vp -> pp wco !vp	1584	vp(pp(从北京到那里)wco(，)!vp(大约有二百多公里))
……			

VP 规则: 1076 条 （宾州树库）

编号	结构规则	频次	示例
4104	VP -> VV NP	23411	VP(VV(涉及)NP(经济、贸易、建设、规划、科技、文教等领域))
4105	VP -> ADVP VP	18157	VP(ADVP(大量)VP(出现))
4108	VP -> PP VP	6012	VP(PP(由上年的百分之三十七)VP(提高到百分之三十九))
4109	VP -> VV VP	5730	VP(VV(没有)VP(发现一例回扣))
4110	VP -> VP PU VP	4813	VP(VP(近年来颁布实行了涉及经济、贸易、建设、规划、科技、文教等领域的七十一件法规性文件)PU(，)VP(确保了浦东开发的有序进行))
4111	VP -> VV IP	4067	VP(VV(防止)IP(出现无序现象))
4112	VP -> VC NP	3493	VP(VC(是)NP(一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程))
4113	VP -> VV AS NP	3120	VP(VV(确保)AS(了)NP(浦东开发的有序进行))
4114	VP -> VP VP	2789	VP(VP(一出现)VP(就被纳入法制轨道))
4115	VP -> ADVP ADVP VP	2641	VP(ADVP(就)ADVP(比较)VP(规范))
4116	VP -> VV PU IP	2546	VP(VV(认为)PU(，)IP(到浦东新区投资办事有章法，讲规矩，利益能得到保障))
4117	VP -> VV QP	2097	VP(VV(增长)QP(百分之四十三点二))
4118	VP -> VV NP IP	2018	VP(VV(使)NP(这些经济活动)IP(一出现就被纳入法制轨道))
4119	VP -> VE NP	1890	VP(VE(有)NP(明确而又具体的规定))
4120	VP -> NP VP	1539	VP(NP(去年)VP(实现))
4121	VP -> MSP VP	1322	VP(MSP(以)VP(确保农牧业生产等重点产业的投入，加大对工业、能源、交通、通信等建设的正常资金供应量))
4122	VP -> ADVP PP VP	1249	VP(ADVP(仍)PP(以轻纺产品)VP(为主))
4123	VP -> PP ADVP VP	917	VP(PP(等积累了经验以后)ADVP(再)VP(制定法规条例))
……			

词库 (lexicon)

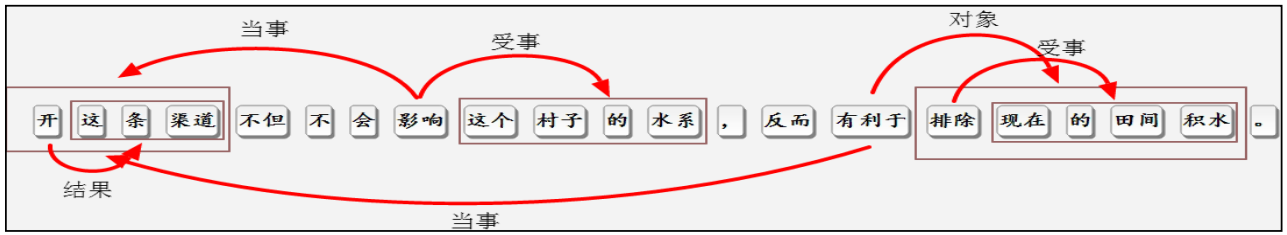
树库（treebank）

命题库（propbank）

篇章关系标注库（discourse treebank）

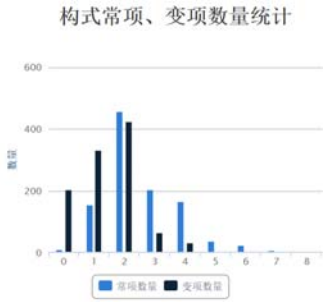
构式库（construct-i-con）

……

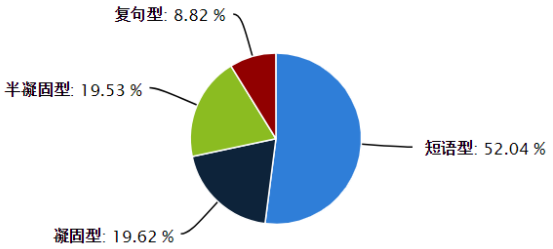


中文深层语义信息标注语料库

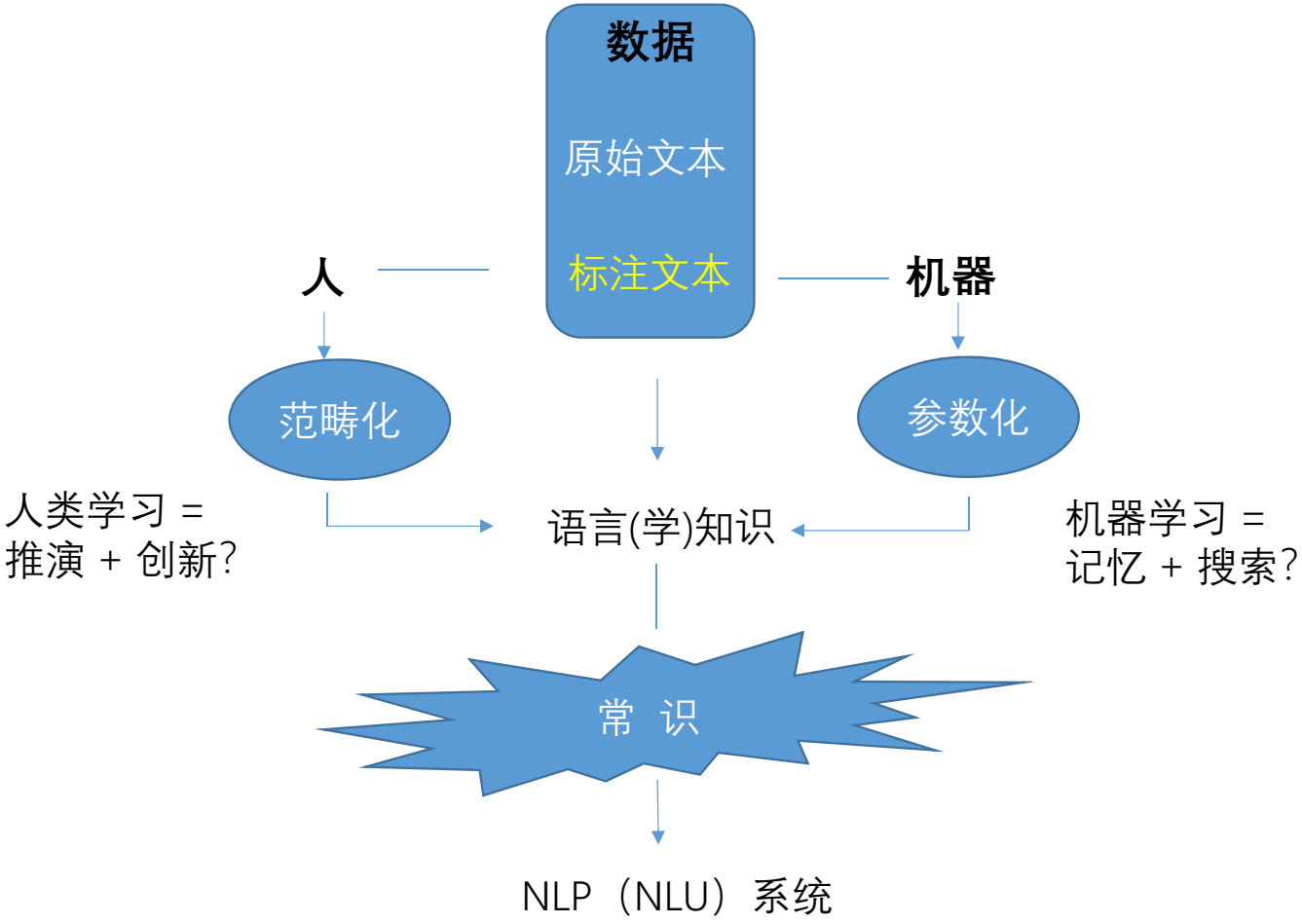
语料类型		规模/字
核心例句集		124074
常用动词例句集		312000
人民日报语料		1131809
微博语料		46259
课题3语料		72964
多领域多文体语料	微信公众号文件	2439183
	小说散文	2004718
	社会科学	2004718
	自然科学	3698172
	中学课本	40446
	科普类	508867
共计		10101324



常项数目	0	1	2	3	4
构式条数	9 (.85%)	154 (14.60%)	456 (43.22%)	204 (19.34%)	165 (15.64%)
常项数目	5	6	7	8	合计
构式条数	36 (3.41%)	23 (2.18%)	7 (.66%)	1 (.09%)	1055



变项数目	0	1	2	3	4
构式条数	205 (19.43%)	330 (31.28%)	423 (40.09%)	65 (6.16%)	32 (3.03%)
变项数目	5	6	7	8	合计
构式条数	0 (.00%)	0 (.00%)	0 (.00%)	0 (.00%)	1055


数据
原始文本
标注文本
人
机器
范畴化
参数化
人类学习 =
推演 + 创新?
语言(学)知识
机器学习 =
记忆 + 搜索?
常 识
NLP（NLU）系统

（分析下面这个句子）

> “美联储主席本·伯南克昨天告诉媒体7000亿美元的救助资金将借给上百家银行、保险公司和汽车公司。” ——《华尔街日报》

在一页纸上，无法画出整个文法分析树 —— 这棵树非常大，非常复杂。应该讲，**单纯基于文法规则的分析器是处理不了上面这么复杂的语句的**。

这里面至少有两个越不过去的坎儿。首先，要想通过文法规则覆盖哪怕20%的真实语句，**文法规则的数量（不包括词性标注的规则）也至少是几万条。语言学家几乎已经是来不及写了。甚至会出现矛盾**，为了解决这些矛盾，还要说明各个规则特定的使用环境。如果想要覆盖50%以上的语句，文法规则的树林最后会多到每增加一个新句子，就要加入一些新的文法。…… 我们即使学了10年的英语语法，也不能涵盖全部的英语。其次，**即使能够写出涵盖所有自然语言现象的语法规则集合，也很难用计算机来解析**。……

自然语言处理的研究也从单纯的句法分析和语义理解，变成了非常贴近实际应用的机器翻译、语音识别、文本到数据库自动生成、数据挖掘和知识的获取，等等

吴军《数学之美》第2版 第2章，自然语言处理 —— 从规则到统计，21-26页

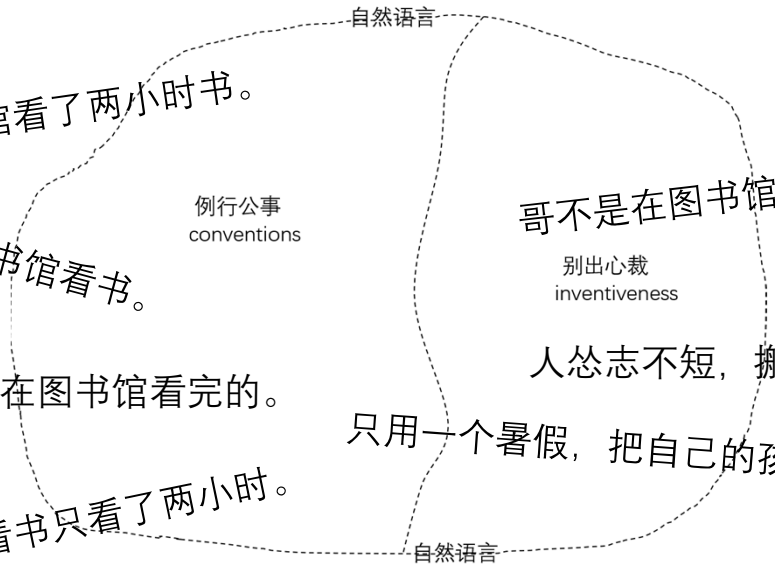
很多支配语言规则的多个观点仍然处于争论中，而其他的规则又看起来过于简单或者过于严格，……

一些深度学习的倡导者争论道：**这种推断的、人为定义的语言学属性是没有必要的，神经网络是可以学习到这些中间表示的（或等价的，或更加好的），这种说法仍然有待裁定**。我当前的个人想法是，在给定充足数据和引导方向正确的前提下，很多的语言学概念确实能够被网络自身学习到。然而，对于很多的其他例子，我们不能获得足够的训练数据，这种情况下，为网络提供更加明确的、清晰的概念将非常有用。即使我们能够获得足够的数据，通过提供更加泛化的概念以及词语的表层信息，我们也想要网络去关注文本或线索的某个方面而忽略其他部分。最后，即使我们不想利用语言学特性作为输入，我们也可能想要使用它们作为补充的监督去指导网络应用于多任务学习中，或者用于设计网络结构，或者用于训练网络范式以更加适合学习某种语言学现象。总的来说，**我们能够看见足够的证据表明语言学概念能够帮助语言理解和系统生成**。

Yoav Goldberg, *Neural Network Methods for Natural Language Processing*, 车万翔 等译，第6章，文本特征构造，65页。

Language is a mixture of regularity and idiosyncrasy.

Ronald Langacker, 1987, *Foundations of Cognitive Grammar*, p.411



我上午在图书馆看了两小时书。

我上午在图书馆看书。

这本书我是上午在图书馆看完的。

我上午在图书馆看书只看了两小时。

我上午在图书馆看书还不到两小时。

……

哥不是在图书馆刷微信，哥刷的是寂寞。

人怂志不短，搬砖忙一晚

只用一个暑假，把自己的孩子培养成“别人家的孩子”!

到了中美贸易战的这一天，我们才发现，一些吹得神乎其神的科技神话、工业神话，真的是神话!

在长达84年20届世界杯的历史上，仅有三支国家队战胜过中国足球队，分别是巴西、土耳其、哥斯达黎加。

—— 文章在哪儿呢?
—— 槽点超标，被404了。
……

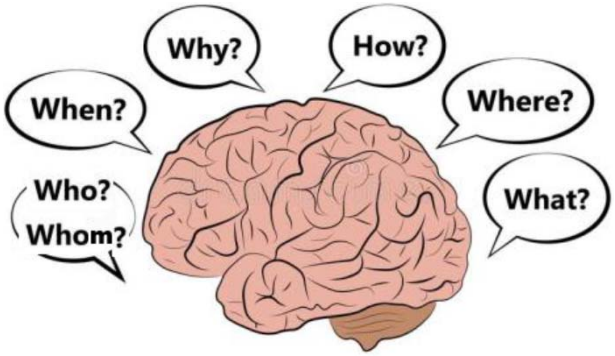
有时例行公事
有时别出心裁
有时一边例行公事，一边别出心裁
你中有我，我中有你，混在一起

自然语言理解:

（1）仅针对例行公事的部分?

（2）兼顾“例行公事”和“别出心裁”?

人和机器，都需要数据! 需要研究**更好的理论**，标注**更多的数据**。



理想工作模式

自然语言
原始数据

人脑 + 电脑

可视化 语言学 **研究平台**

标注工具 | 统计工具 | 分析工具 | 调试工具 | 测试工具 | 管理工具

语言(学)知识
[描写 + 解释]

范畴化的 | 数据化的 | 可视化的 | 多层级的 | 跨语言的 | 多粒度的 | 可更新的 | 易维护的

一堆“**字符串**”
00101001110…，
哦！不！我看到了
“人类**语言**”！！！

谢 谢!

zwd@pku.edu.cn